4.0 SAMPLING METHODS

4.1 Introduction

One of the weaknesses in ecology today is that too many investigators fail to realize the importance of sampling. A logical reason for this difficulty is that studies are often centered on one or two study areas, so that the investigator tends to forget that he is in fact studying a sample from some larger population. This may not be seen as a handicap until it becomes necessary to attempt to extend the results of the study to the larger area. The perceptive observer may then suddenly realize that there really isn't much basis for such an extrapolation, unless he does in fact have data from a number of subareas (i.e., a representative sample) on which to base the extrapolation and to provide a basis for assessing its validity.

The intent here is to provide a brief overview of sampling methodology. Most of the material follows the lines of survey sampling methods, as given in much more detail by Cochran (1977). Thompson (1992) includes methods of particular interest in ecology. The very basic features are those of an elementary statistics course. Most students will prefer to refer to familiar textbooks for these aspects. The essential material has to do with some elementary statistical concepts and a few standard distributions, mainly the binomial, hypergeometric, Poisson, and normal. Students not familiar with these distributions and the basic rules of probability should look them up in one of the elementary references. A brief sketch of the statistical background appears also in Chapter 1.

4.2 Simple random sampling

The main complication in defining simple random sampling is one of defining the meaning of the word "random". Our approach is that of probability theory, in which it is assumed that every sampling unit (some sort of explicitly defined entity) has the same probability (chance) of being drawn into the sample. The mechanics of drawing a random sample then depend on giving each unit the same chance of inclusion in a sample while keeping the choices independent of one another. The standard procedure is to assign a number to each unit in the population, and to refer to a table of random numbers as a device for selecting the sample.

Once the sample has been drawn and measurements made on the sample units, various problems of analysis of the data must be dealt with. However, procedures for analysis of the data need to be considered well in advance of the sampling to be sure that the right kinds of data are collected. That is, the investigator must first prepare a sampling plan, which designates exactly how the sample will be obtained. Secondly, there should be a definite plan for the analysis of the resulting data, specifying what statistical analyses will be carried out, and what will be done if a particular kind of result is obtained in the analysis. Many of the problems in field research are caused by the lack of such a study design. It may be objected that one cannot produce such a plan if it is not known in advance how the study will turn out. There are several answers to this objection. One is that few studies are conducted in completely new situations. Usually there are previous investigations that can be used in the design stages, and data that can be used to test both the sampling plan and the analytical procedures.

4.3 The finite population correction

The estimated variance of a mean for sampling without replacement from a finite population is:

$$s^2(\bar{x}) = \frac{s^2}{n} \left[\frac{N - n}{N}\right] = \frac{s^2}{n} \left[1 - \frac{n}{N} \right] \qquad (4.1)$$

and the finite population correction is just the quantity in brackets, or one minus the fraction of the population actually sampled (frequently designated as f). Thus when nearly all of the population is taken into the sample, the variance of the estimated mean becomes very small, as it logically should.

When the fraction of the population sampled is small, this equation implies that size of the population has little effect on the standard error and thus on confidence limits. This is a result that comes as a surprise to many people, who intuitively suppose that bigger samples are required for very large populations. This is, however, simply not true. A large population may, of course, offer more logistic problems in sampling and thus be more expensive to sample.

As a general rule of thumb, when the sampling fraction is less than about 5%, it is customary to neglect the finite population correction factor, and treat the sample as though it had been obtained by sampling with replacement. Sampling with replacement refers to circumstances where objects can be drawn from the population one at a time and replaced before the next object is drawn. With such a process, probabilities remain unchanged as the drawing proceeds, making calculations much simpler than if replacement does not occur, when removal of one individual changes the odds on selecting others in the next draw.

Students whose statistical training has come from courses in which hypothesis testing was mainly emphasized may not have encountered the notion of a finite population correction. This is because most tests of hypotheses are formulated on the basis of sampling from an infinitely large population, or on the basis of sampling with replacement.

4.4 Confidence intervals

Ideally, one would like to be able to know how far "off" a particular estimate is from the true parameter value. Statistical methods offer no such utopian result, and the best that we can do is to make probability statements that apply to the long-run of future trials, or to some hypothetical population just like the one currently under study. These take the form of confidence limits, which are a statement of the following kind:

$$P_r\{X_L \leq \mu \leq X_U\} = 1 - \alpha \qquad (4.2)$$

where X_L denotes the lower confidence limit, X_U the upper, and the probability that the true, but unknown, value (μ) of the random variable of interest will fall between these limits is 1 - α. The proper interpretation of this statement is that a very large number of repetitions of the "experiment" at hand would yield confidence limits that include the true, but unknown, μ in a fraction 1 - α of the trials. It must be emphasized that the statement cannot be interpreted as pertaining to a single set of sampling results that are in hand. Such a statement would be ridiculous because the confidence interval then either includes the true value or it does not, and no probability is involved--it's just that we have no way of knowing where the true value lies. Hence we have to adopt some sort of long-run view of the "odds on being right."

One of the most common mistakes in reporting the results of a statistical analysis is to assert that "the probability is l - α that the hypothesis is false". Just as with the confidence limits statement above, a testable hypothesis is either true or false, but there is no need for statistical analysis if one knows the answer. If the answer is not known, then the statistical approach attempts to supply some quantitative assessment of the "odds" for and against the hypothesis. The problem that many people have with this is that they have been admonished from childhood to "make up your mind". Such decisions should be stated as a belief based on the evidence, but announced separately from the probability statements used to assess the evidence.

Most investigators tend to use confidence limits that are symmetric about the estimate. No doubt this is a consequence of the common use of the symmetric normal distribution, which leads one to tend to cut off about α/2 of the probability distribution on each side of the mean, and thus get symmetric limits. In point of fact there is nothing in theory or practice that says that the limits should be symmetric--all that is required is that there be 1 - α of the distribution within the limits. Also, setting limits for a distribution like the Poisson is likely to result in asymmetric limits. One reason is the difficulty of cutting off an exact fraction (α) of the distribution when one must set the limits in terms of integer values. This difficulty can quickly be appreciated by trying some examples with tables of the Poisson distribution.

For ease in understanding and remembering the procedure for obtaining confidence limits, we will use the standardized or unit normal curve, and reverse the usual process of going from some other normal distribution to the standardized--that is we now look up a value (z) in tables of the unit normal that cuts off the desired proportion, α, of the distribution. If α is to be 0.05, then we find z = l.96, and

$$\frac{X_c - \mu}{\frac{\sigma}{\sqrt{n}}} = \pm 1.96 \tag{4.3}$$

where X_C represents upper or lower confidence limit respectively corresponding with the plus and minus signs on the right hand side of the equation. Thus we have $X_C = \mu \pm 1.96\ \sigma/\sqrt{n}$ and the probability statement previously given is satisfied by the corresponding choices of X_C (which are X_L and X_U). In practice, it is necessary to substitute $\bar{x}$ for μ.

The above results assume that one knows σ, which is almost always not the case. If the sample size is about 30 or more, it really doesn't matter much, in that $s/\sqrt{n}$ then provides an adequate estimate of the standard error ($\sigma/\sqrt{n}$). If sample size is small, then it is preferable to use tables of the "t" distribution (instead of the unit normal distribution), which make allowance for uncertainty about the variance. Values of the t-distribution are available in EXCEL.

One also needs to bear in mind that the estimate of σ (usually denoted by s) that is appropriate depends on what quantity one is setting limits on. For a mean, we proceed as above, but if limits on a single observation are to be secured, then we naturally use the standard deviation, s, in place of $s/\sqrt{n}$ ·

Example 4.1 Calculating confidence limits

For "95 percent confidence limits" ($\alpha = 0.05$), many people round 1.96 to 2.0 so that the limits can be calculated from $x_C = \bar{x} \pm 2s/\sqrt{n}$. Suppose s = 9, n = 16 and $\bar{x}$ =10. Then we have $x_C = 10 \pm 2(9)/4$, which can be expressed as $5.5 \leq \mu \leq 14.5$.

4.5 Determining sample size

Determining the sample size required to provide confidence limits of preassigned width on a mean is again a matter of using the z values. From the results above we have:

$$\pm \frac{zs}{\sqrt{n}} = X_C - \mu \quad \text{and we let } D = |\mu - X_C|$$

where the vertical lines denote "absolute value of...", so that D amounts to the half-width of the desired (symmetric) interval. Hence:

$$n = (zs/D)^2 \qquad (4.4)$$

Another way to approach sample size estimation is to express D relative to the mean:

$$D/\bar{x} = \frac{zs}{\sqrt{n}\ \bar{x}}$$

so that

$$n = [z(c.v.)/D(\%)]^2 \qquad (4.5)$$

where c.v.= $s/\bar{x}$ is the coefficient of variation, and D(%) = $D/\bar{x}$ expresses D as a proportion (note that D(%) is used as a proportion, NOT as a percentage). The utility of this approach is that we often have an idea of the coefficient of variation, but may not know what the mean is likely to be, so it is possible to set proportionate limits this way. The advance specification can then be, for example, that "I want 95% confidence limits no wider than ± 20 percent of the mean".

When sampling without replacement is important (i.e., when the sample is likely to be a significant proportion of the population,say, 5% or more) the above relationships serve to get an approximate value of the sample size needed, which can be labelled n_o. The final estimate of n is then obtained from:

$$n = \frac{n_o}{1 + \frac{n_o}{N}} \tag{4.6}$$

This estimate results from using a variance estimate that includes the finite population coefficient as previously described (Equation 4.l).

Example 4.2 Determining sample size

Suppose we want smaller (narrower) confidence limits, say $8 \leq \mu \leq 12$, in Example 4.1. Using Eq.(4.4) D = 2, and $n = [2(9)/2]^2 = 81$. If an appreciable fraction of the population is to be sampled, then the above result needs to be corrected by using Eq. (4.6). Assume N = 100, then we have n = = 81/(1 + 81/100) = 44.8, which may be rounded to n = 45. Suppose it is required that D(%) = 0.1 for an approximate 95% confidence interval of ± 20%. Then $D(\%)=0.1=2(9)/10\sqrt{n}$, and n= 324, which exceeds the supposed population size of N=100. But Eq.(4.6) gives n=76.

4.6 Stratified random sampling

Almost invariably, ecologists have some advance information about populations that they wish to sample. This prior knowledge may well be one of the reasons for rejecting random sampling and substituting some sort of purposive selection, wherein one chooses sampling units that "look" tobe representative or typical. There are, however, methods that take into account advance knowledge and at the same time provide the protection against bias that is given by random selection. One such method is to classify all of the population units into one of several strata (groups), and to then take random and independent samples in each such stratum.

The name, stratified sampling, comes from the close analogy to the layering effect seen in various circumstances, since we normally attempt to have the strata represent gradations in value of the random variable of interest. If one can do a fairly good job of segregating units by magnitude of the random variable under consideration, then it is apparent that the variability to be encountered in sampling within a particular stratum may be substantially reduced over that without stratification. Hence greater overall precision results for a particular total sample size.

Stratified sampling will require somewhat more advance effort than simple random sampling, but the usual result is that it turns out to be less extra effort than one might suppose. The method provides some side benefits in terms of better understanding of the material being studied, and of the nature of the problem dealt with. In some circumstances it may be that a portion of the sampling units in the population are very difficult to reach, or otherwise expensive to sample. In this case, stratification can specifically take differences in cost into account, and provide alternative sampling schemes aimed at getting the most information for the effort expended.

The material here is devoted to an exposition of the basic technology in stratified sampling. Many additional features will be found in texts like that of Cochran (l977) and Thompson (1992). Here we will deal with such things as how to go about stratifying a population, the estimation procedures, various equations for obtaining variance estimates, and determining sample size.

4.7 Mechanics of stratification

Undoubtedly the most common ecological application of stratified sampling has to do with locating sampling plots or other measurement schemes on a map of a region of interest. The process of stratification is then intuitively obvious--one finds a way to break the map up into strata, each of which is composed of some large number of individual sampling units, usually plots of square or rectangular shape. It might be noted that the units in one particular stratum do not have to be contiguous--this is in fact one of the primary advantages of stratification. However, it is advantageous to keep units in a stratum more or less contiguous if the survey is designed for analytical purposes (e.g., to make comparisons between strata).

The basic process is just to assign units to strata according to the available prior information, seeking to get the units in any one stratum as much alike as possible in terms of the random variable or variables being studied. The next step is to assign serial numbers to every unit in each stratum. This does not, of course, require that someone write down all the numbers--all that is required is a trustworthy scheme for assigning and finding again any particular number. Often it will turn out to be simplest to delineate stratum boundaries with colored pencil and to note the starting and ending point of each row of units in a given section of one stratum by writing the corresponding numbers on the map. Sometimes large blocks can be counted as the equivalent number of sampling units, that is sampling units might be mil-acre (or meter-square) plots but the strata may be made up of larger units. A little practice soon settles the details for any particular set of conditions.

We use the following notation, which follows that of Cochran (l977) for convenience in referring to the much more complete description available there.

N_h $(h = 1, 2,..., L)$ The number of units in the stratum. There are L strata in all, and $N = N_1 + N_2 + ... + N_L$

n_h The number of units in the h^{th} stratum that are selected for enumeration (a random sample of n_h units from the h^{th} stratum). In most circumstances n_h should be at least 4.

y_{hi} $(i = 1, 2, ..., n_h)$ The observed value of the random variable (Y) on the i^{th} unit of the h^{th} stratum. There are n_h such observations taken from the h^{th} stratum.

$W_h = \frac{N_h}{N}$ The proportion of the total units present in the h^{th} stratum (N units in population; $N_1 + N_2 + ... + N_L = N$).

$\bar{y}_h = \Sigma y_{hi} / n_h$ Average of the sample units from the h^{th} stratum.

$f_h = \frac{n_h}{N_h}$ The sampling fraction in the h^{th} stratum (i.e., the fraction of the units in the h^{th} stratum that are actually examined.

4.8 Estimates from a stratified sample

Since the W_h represent the proportion of the total population in the h^{th} stratum, they are logical weighting factors for estimating the overall population mean. The equation is:

$$\bar{y}_{st} = w_1\bar{y}_1 + w_2\bar{y}_2 + ... w_L \bar{y}_L \tag{4.7}$$

where $\bar{y}_{st}$ refers to the mean of a stratified sample.

By the rule for variance of a linear combination of independent random variables (independent because of the random sampling in separate strata), we have the variance of the estimate as:

$$V(\bar{y}_{st}) = w_1{}^2V(\bar{y}_1) + w_2{}^2V(\bar{y}_2) + ... + w_L{}^2V(\bar{y}_L) \tag{4.8}$$

This result assumes that the sampling fraction (f_h) in each stratum is small enough to neglect the finite population correction. If the fpc is included, the equation for variance in the h^{th} stratum is:

$$V(\bar{y}_h) = \frac{S^2{}_h}{n_h}\frac{(N_h - n_h)}{N_h} = \frac{S^2{}_h}{n_h}[1 - f_h] \tag{4.9}$$

where $S^2{}_h$ is obtained from:

$$S^2{}_h = \sum_{i=1}^{N_h}(y_{hi} - \bar{y})^2 / (N_h - 1) \tag{4.10}$$

These two equations can be combined to get the desired result:

$$V(\bar{y}_{st}) = \sum_{h=1}^{L}(W_h{}^2S_h{}^2/n_h) - \sum_{h=1}^{L}W_hS_h{}^2/N \qquad (4.11)$$

However, S_h^2 is not a quantity that can be determined from sampling, being the variance of the entire population in the h^{th} stratum. A logical procedure is to estimate it from:

$$s_h{}^2 = \sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2/(n_h - 1) \qquad (4.12)$$

which is the usual variance estimate for a simple random sample. The last term in the equation for $V(\bar{y}_{st})$ is the finite population correction so if all of the f_h are small (say, less than 5%) this term can be dropped

4.9 Confidence limits

When sample sizes in the several strata are all substantial, the confidence interval takes on the same form as we have previously encountered it for simple random sampling, and can be written as:

$$\bar{y}_{st} \pm z\, s(\bar{y}_{st})$$

where z is the value from the unit normal curve corresponding to the confidence level wanted, for example z = 1.96 for $\alpha = 0.05$. We now use $s^2(\bar{y}_{st})$ or, in this case its square root, to denote that this is an estimate of the true variance given by Equation (4.11).

When sample sizes for individual strata are small, as is not uncommonly the case, there is a difficulty brought in by the fact that use of z corresponds to virtually knowing the true variance. As was noted earlier, samples of 30 or so give close enough estimates that one does not need be too concerned about an effect on the confidence limits. With only a few observations in one or more strata, however, the estimates of stratum variance (s_h^2) may not be so precise, and this situation would usually be handled by substituting a "t" value for the z value, that is, by making use of the t-distribution, which allows for uncertainty about the variance estimate. The trouble here is that we need to combine the several stratum variances to estimate $V(\bar{y}_{st})$, but it is not proper to average the various "t" values corresponding to stratum sample sizes (we would ordinarily look up a t-value with n_h -1 degrees of freedom for each stratum). Cochran (l977:96) and Thompson (1992:106) give a rather complicated expression for calculating an "effective" number of degrees of freedom for use in this case.

Example 4.3 A caribou census

An example that incorporates nearly all of the basic elements in stratified random sampling is furnished by an aerial census of Alaskan caribou. Details will be found in a paper by Siniff and Skoog (1964). We here extract that part which is appropriate for illustration. Six strata were selected on the basis of preliminary observations of relative caribou abundance, and delineated on detailed maps. Sampling units were four-square-mile blocks, and each such unit within each stratum was assigned a number. There were no advance estimates of within-stratum variances available, so the stratum standard deviations (S_h) were assumed proportional to the preliminary rough estimates of population levels in the stratum. This provided the following data for allocation according to Equation (4.14):

Stratum	N_h	W_h	S_h	$\frac{W_h s_h}{\Sigma W_h s_h}$	n_h(opt.)	n_h (actual)
A	400	.572	3000	.428	96	98
B	30	.043	2000	.022	5	10
C	61	.087	9000	.195	44	37
D	18	.026	2000	.013	3	6
E	70	.100	12000	.299	67	39
F	120	.172	1000	.043	10	21
	699	1.000	29000	1.000	225	211

The "optimum" allocation was based on the total number of sample units (225) that the investigators believed could be surveyed in the time available. The actual allocation amounted to "hedging" against uncertainty about the likely values of S_h. Thus there were several strata (B, D and F) where the supposed optimum allocation called for rather small samples, so these were increased at the expense of strata (C and E) where the optimum plan called for censusing a substantial fraction of the units in the stratum. Survey results were as follows (N_h, W_h were as used above, and n_h as given in the last column above):

Stratum	$\bar{y}_h$	s_h^2	$\frac{W_h^2 s_h^2}{n_h}$	$W_h s_h^2$	$\bar{y}_h W_h$
A	24.1	5,575	18.613	3,189	13.79
B	25.6	4,064	.751	175	1.10
C	267.6	347,556	71.098	30,237	23.28
D	179.0	22,798	2.569	593	4.65
E	293.7	123,578	31.687	12,358	29.37
F	33.2	9,795	13.800	1,685	5.71
	-	-	138.518	48,237	77.90

From Equation (4.7), $\bar{y}_{st}$ is the sum of the last column above, or 77.9 caribou per four-square-mile unit. This is readily converted to a total for the area surveyed by multiplying by the total number of units, giving (77.9)(699) = 54,450 caribou.The variance estimate (Equation (4.11) is:

$$v(\bar{y}_{st}) = \sum_{h=1}^{L} (W_h^2 S_h^2)/n_h - \sum_{h=1}^{L} W_h S_h^2/N = 138.518 - (48{,}237/699) = 69.51$$

Confidence limits on the mean may be obtained by assuming z to be equal to 2 (or 1.96 to be exact, for $\alpha = 0.05$), since rather substantial samples are involved here, in most strata, giving:

$\bar{y}_{st} = \pm 2(69.51)^{1/2}$ or 77.9 ± 16.7 caribou per four-square-mile unit.

For total caribou on the study area, we estimate the variance as:

$v(y_{tot}) = N^2 v(\bar{y}_{st}) = (699)^2(69.51) = 33{,}962{,}655$

and limits are:

$54{,}540 \pm 2(33{,}962{,}655)^{1/2}$ or $42{,}885 \leq x_{tot} \leq 66{,}195$.

Notice that s_h^2 increases with increasing y_h in the table above. The investigators plotted $\log_{10} (s_h^2)$ against $\log_{10} \bar{y}_h$ (Siniff and Skoog, 1964:398) and obtained a regression relationship:

$$y = 1.63 + 1.42\ x$$

where $y = \log_{10} (s_h^2)$ and $x = \log_{10} \bar{y}_h$. This is equivalent to the relationship:

$$s_h^2 = 42.66(\bar{y}_h)^{1.42}$$

which might be used to estimate variances in planning similar surveys. However, it is important to remember that size of the sampling unit (4 sq. mi. in this case) will affect such a relationship.

Example 4.4 A mortality survey

A Michigan study of over-winter losses of whitetailed deer (Whitlock and Eberhardt, 1956) provides an example where the finite population correction is negligible. In this case, nearly 19,000 square miles (all of the northern lower peninsula of Michigan) were classified into five strata on the basis of estimates made by field biologists. The primary units were half-sections (one-half square mile), but these were subsampled in the actual search by using a strip 88 yards wide laid out as a rectangular course 1/2 mile long and 1/4 mile wide. Width of the strip was based on use of four-man teams, with each individual responsible for searching a 22 yard wide interval. Various complications were involved in the design inasmuch as it was necessary to consider prospects for missing dead deer on the strip, the number of men available in various locations (and transportation), the necessity for one man to act as compass-man, need for a biologist in each crew, and so on.

Advance data from a previous survey of an area of high mortality suggested that the coefficient of variation ($s/\bar{x}$) might be about 1.30, so estimates of S_h were obtained by multiplying 1.3 times an estimated number of deer to be found on each plot (these guesses were made in the process of setting up strata).

It was decided that about 110 plots could be surveyed with available manpower so the allocation was devised as follows:

Stratum	Expected losses per sq. mile	Expected losses expressed as dead deer/plot	Square miles in stratum	W_h	Estimate of S_h	Preliminary allocation
I	20+	3.75	408	.0220	4.88	23
II	10-20	1.88	1,048	.0566	2.44	29
III	5-10	.94	2,293.5	.1240	1.22	32
IV	1-5	.31	5,567.5	.3010	.40	25
V	0-1	.01	9,181.5	.4964	.01	1
			18,498.5	1.0000		110

The allocation again followed Equation (4.14) but four plots were added to stratum V, giving 114 in all, of which 113 were actually searched (one plot was completely flooded at the timeof the survey). Survey results were:

Stratum	n_h	s_h	$\bar{y}_h$	Contribution to to $V(\bar{y}_{st})$	Coefficient of variation	$s_h^2/\bar{y}_h$
I	23	2.146	1.826	.000097	1.18	2.52
II	29	1.082	.621	.000129	1.74	1.88
III	31	.724	.484	.000260	1.50	1.08
IV	25	.541	.280	.001062	1.93	1.04
V	5	--	0.00	.00	--	--
	113			.001548		

In this instance, only very small fractions of each stratum were searched so Equation (4.11) reduces (by dropping the right-hand term) to:

$$v(\bar{y}_{st}) = \sum_{h=1}^{5} (W_h^2 S_h^2)/n_h$$

and the individual terms are listed under the heading "Contribution to $v(\bar{y}_{st})$" so that one can see in which stratum most of the variability turns up. Comparing the expected losses and the $\bar{y}_h$, it becomes apparent that the over-estimates were largely in strata I and II, which was not especially surprising since the winter turned out to be milder than anticipated when the survey was planned, and starvation losses were correspondingly lower (major starvation areas nearly all were in strata I and II).

The coefficients of variation in the above table show the advance estimate to be somewhat low.The last column of the table gives $s_h^2/\bar{y}_h$

which is the "index of dispersion" and is unity (within sampling or "chance" errors) for a Poisson distribution. This suggests that such a distribution (i.e., wholly random dispersal of dead deer) may apply to strata III and IV, in which case the S_h for allocation might simply have been taken as equal to the square roots of the expected numbers of deer per plot.

4.10 Allocating the sample to strata

We have thus far gotten the cart before the horse, having considered how to analyze sample results without considering how the total sample ought to be distributed over the strata. Two kinds of allocation are in common use. The first is perhaps what one would expect to do without any advance information about the variability in various strata, that is, distribute the sample in proportion to the size of the strata ("proportional allocation"). This is also known as a self-weighting sample, since fractions going into each stratum will be equal to W_h, so that a simple mean of all of the sample results will be equal to the weighted mean previously given. In this case we have $n_h/N_h = f = n/N$ so that the sampling fraction is the same in all strata.This leads to a simpler expression for the variance:

$$V(\bar{y}_{st}) = \frac{1 - f}{n} \sum_{h=1}^{L} W_h S_h^2 \qquad (4.13)$$

and we again have to substitute sample estimates for S_h^2 .

Proportional allocation is easy to accomplish and to analyze, but often is not a very efficient way to use sampling resources. In most ecological work it turns out that the variance and mean tend to increase together, so that the strata likely will have rather different variances, and proportional allocation will then undersample some strata and oversample others. An allocation which allows for the effect of differences in stratum variances is the scheme called "optimal allocation". This method can be shown to minimize $V(\bar{y}_{st})$ for a fixed n. Optimum allocation is given by the following relationship:

$$n_h = nW_hS_h / \sum_{h=1}^{L} W_hS_h \qquad (4.14)$$

Of course use of the formula demands at least a guess at the S_h. In many studies, there will be some preliminary information about the magnitude of variances to be encountered, quite often in the form of coefficients of variation, which may be applied to the expected stratum means to get an estimate of stratum standard deviations. It also turns out that this kind of allocation is not too sensitive to errors in advance estimates of S_h so one can usually expect to do a better job with this method so long as the stratum variances do differ appreciably and the guessed values of S_h are in the right "ballpark".In many natural populations the stratum with the lowest mean can be expected to have roughly a Poisson distribution of individuals (assuming the purpose of the survey is to estimate total individuals) so the investigator can set that variance equal to the expected mean density, and go on from there on the basis of any information about how variability increases with the means of the remaining strata.

One further feature of allocation worth considering here is the circumstance where sampling costs differ among strata. Perhaps the simplest assumption is that total cost of thesampling can be written as:

$$\text{cost} = c_o + \sum_{h=1}^{L} c_h n_h \qquad (4.15)$$

where c_h represents the cost of measuring each sample unit in stratum h, the c_h are not all equal, and c_o is a fixed or "overhead" cost. In this case, Cochran (l977:97) shows that the allocation should be:

$$n_h = \frac{n W_h S_h}{\sqrt{c_h}} / \sum_{h=1}^{L} W_h S_h / \sqrt{c_h} \qquad (4.16)$$

so that the number of samples in a stratum depends on the stratum size, its variability, and cost of sampling. One takes more samples in large and variable strata, but also increases sample size if sampling is cheap in the stratum.This kind of allocation can be rather useful in dealing with sampling problems where either access or measurement may be quite difficult for part of the population. It is worth noting that other kinds of cost functions might be obtained from knowledge of the sampling problem, and special allocations then devised. Cochran (l977) discusses "cost functions" for various circumstances.

Example 4.5 A deer population estimate

Counts of "pellet-groups" have been used to estimate deer populations for many years. Daily defecation rates are remarkably constant (about 13 groups per day) and over-winter accumulations of pellets can be identified by the underlying mat of leaves dropped the previous fall. There is thus a straightforward conversion from numbers of pellet-groups to "deer-days" which in turn can be converted to average population levels for the over-winter season. Stratified random sampling has been used to conduct such surveys in northern Michigan for more than 25 years. About 35,000 square miles are surveyed, requiring on the order of 500 man-days of effort. Some nine separate areas (Game Management Districts) are surveyed independently. An example for one such are (District 7 in 1962; Ryel 1971:131) appears in the following table:

Stratum	Area (sq.mi.)	Prop.(W_h)	n_h	$\bar{y}_h$	$\frac{W_h^2 s_h^2}{n_h}$
I	190	.0541	9	65.22	1.7568
II	425	.1211	12	29.25	1.2649
III	1544	.4399	34	15.35	1.7803
IV	1144	.3259	10	10.70	1.7748
V	207	.0590	1	0.0	--
	3510	1.0000	66	-	6.5768.

The overall weighted mean number of groups per sampling unit was

$\bar{y}_{st}$ = 17.31 with two standard errors being 29.6 percent of that estimate. With the very large areas involved, an appreciable amount of time is expended in traveling to the sampling units. Since experience shows that the individual plots should not be very large (to avoid missing groups in the counts), a cluster sample is used at each site, comprised of eight individual plots arranged along a half-mile line. For convenience, square miles (sections) serve as the primary sampling unit, with a random starting place and distance from the boundary used to locate each systematically arranged cluster of plots.

To reduce the effort required to plan and execute these large-scale surveys (150 to 200 people may be involved annually), the same plots have been used for a number of years in succession. This makes it possible for the field men to plan their work efficiently, since they know the plot location well in advance and can anticipate just when the plots will be accessible (and free of snow). Ryel (1971:222) calculated the optimum allocation for a number of years. Results for the District used as an illustration above are:

Stratum	Actual allocation	Calculated optimum allocation					
		1959	1960	1961	1962	1963	1964
I	9	18	14	6	13	4	6
II	12	11	14	21	13	19	18
III	34	27	29	21	26	32	32
IV	10	9	9	18	14	11	10

It can thus be seen that the original allocation was, on the average, quite satisfactory. Two kinds of factors may affect these results. One is that the distribution of deer may change somewhat from year to year, in consequence of winter weather conditions. Another is that variances for each stratum are estimates, and thus will vary somewhat due to chance alone.

4.11 Further remarks on stratified sampling

Cochran (1977), Thompson (1992), and other texts on survey sampling supply a good deal of auxiliary information on methods and techniques for various special cases. A few points that are examined in more detail in those references are summarized here:

(1) Gains in stratified sampling for the estimation of a proportion are usually not sizable unless the proportion (P) varies sharply from stratum to stratum, and in most cases, proportional allocation is preferable.

(2) Many surveys are designed to measure more than one random variable, whereupon the question of allocation gets complicated. An initial approach is to calculate allocation for the variables of main interest separately and determine whether the several allocations differ appreciably. If so, then it may be possible to devise some sort of cost function to help in a decision. If

there are two variables of main interest, then there are some handy schemes for handling the two in a single allocation.

(3) We have not considered how to determine either the number of strata nor stratum boundaries. Most studies of natural populations will involve the location of strata on maps, and the quantity of advance information on the population will usually be such that the number of strata probably will not be less than three nor more than six. A few trial efforts at laying out strata will usually resolve most questions of boundaries. Probably the major difficulty comes up if large areas are to be covered, so that there are a number of people involved, each having rather good local knowledge. Then the principal job turns out to be in getting individuals to agree on what constitute definite strata, and how they should match at the junction of two districts where different people are locally "expert" on the subject matter. Some one person usually has to umpire the decisions, and this can perhaps be done after individuals have made up maps reflecting their knowledge.

(4) Sometimes it is possible to make use of stratification after a simple random sample has been taken. It must be emphasized that stratification cannot legitimately be undertaken on the basis of examining the sample results, but it may turn out that it is not possible to assign individual units to strata until after they have been surveyed, that is, the total number of units in each stratum may be known in advance, but the stratum to which a particular sample unit belongs cannot be determined until the measurement is made.

(5) Most experience with natural populations shows that variability increases with the mean. This is fairly sound grounds for recommending that "optimum" allocation always be carefully considered before selecting one of the other possibilities.

4.12 Ratio Estimation

The main results for ratio estimation require that the population total of an auxiliary variate (X) be known, and the correlation between X and the variable of main interest (Y) is used along with the known total to obtain an estimate of either the mean of Y or its total with greater precision than may be obtained from simple random sampling of Y alone. So far, ratio and regression methods have been little used in ecology and resource management surveys, partially perhaps because of a lack of suitable correlated variables with known totals, but also because many investigators are not familiar with these methods.

In the usual notation, X is used to represent the known population total. Since we have been using X to represent a random variable, X_T will denote the population total here. The ratio estimate is:

$$\hat{Y}_R = \frac{\Sigma y_i}{\Sigma x_i} X_T = \frac{\bar{y}}{\bar{x}} X_T$$

as an estimator of the population total for Y. The mean value of Y is estimated by replacing the population total (X_T) by the mean above.

The population ratio is estimated by:

$$\hat{R} = \frac{\Sigma y_i}{\Sigma x_i}$$

If interest is principally in the population ratio, then it is not necessary to know X_T.

An important application of ratio methods is worth mentioning here in order to provide an illustration of the nature of the above relationships. This is the use of "strip transects" (discussed in more detail in Chap.5) on irregularly-shaped areas. A strip transect is just a long, narrow plot extending completely across a study area. For present purposes, we assume that all of the objects of interest are counted on each of a number of strips. Each such transect then constitutes a sample plot. If the region under study is rectangular in shape, then each sample plot will have the same area, and no adjustment is for transect length is needed. However, in most practical situations, study areas will be irregular in shape. Strip transects across such a site will thus have different individual areas, presenting a problem in the analysis of the data, since plot size is now also a random variable.

It is true that a simple random sample of strips will provide an unbiased estimate of the total number of objects on the study area. The appropriate random variable is the total number on each strip, and the calculations proceed as previously described for simple random sampling of a finite population (the total number of possible sample strips). However, such a theoretically correct result is of almost no practical interest in dealing with natural populations, just because such populations exhibit high variability even with efficient methods of sampling. We thus cannot afford to bring in any further variability. Ratio methods can conveniently be used to resolve the problem simply by regarding the area of each sample strip as X_i, so that X_T is the total area of the study region, and letting Y_i represent the total number of objects on each sample plot. We then have that $\hat{R}$ estimates the average density (number per unit area) observed in the sample, and $\hat{Y}_R$ is an estimate of the total number of objects on the entire study area.

The ratio estimate is biased, but the bias is considered unimportant for large samples. In this case, a rule of thumb is n of at least 30, and the coefficients of variation of the means of X and Y should both be less than 0.10 (Cochran, 1977:153). Stratification and ratio estimation may serve roughly the same purposes, and it is likely that an effective stratification could be obtained through the use of the auxiliary variable X. Thus in the example given above, one could stratify the study area into blocks such that the length of potential sample transect strips is about constant in each stratum. However, the ratio method provides a "natural" approach in this instance, and is thus the appropriate choice.

The ratio estimate effectively assumes the relationship between Y and X to be Y = RX + e, where e represents an "error" component and R is an unknown constant. In some instances it may not be reasonable to assume that the relationship goes through the origin, so that a regression estimate is appropriate. This method is also biased, so that large samples are generally recommended. Details are available in Cochran (1977) , Thompson (1992), and many other texts on sampling. Before undertaking to use the ratio or regression techniques, an investigator should have some preliminary observations (or good general knowledge) that indicate a relationship between the variable of primary interest (y) and an auxiliary variable (x). A first step it plot the data and to note whether the regression line clearly does not go through the origin. If this is the case, then it is advisable to look into regression estimation, rather than using ratio methods. Occasionally there may be more than one useful auxiliary variable and it is then possible to use multiple regression.

4.13 Variance of ratio estimates

An estimate of the variance of a ratio is given by:

$$V(\hat{R}) \doteq \frac{1 - f}{n\bar{X}^2} [\sum_{i=1}^{N} (Y_i - \hat{R} X_i)^2/(N-1)] \qquad (4.17)$$

Here f again represents the finite population correction, and may be neglected if n/N is less than about 5 to 10 percent. N is the total number of units in the population, n, the number in the sample, and $\bar{X}$ the population mean of the auxiliary variate. Note that the summation runs over the entire population, so that this is an approximation to the "true" variance, and it will in turn have to be estimated by a quantity that can be calculated from a sample; that is, we replace the quantity in the right-hand brackets by sample data, getting:

$$s^2(\hat{R}) \doteq \frac{1 - f}{n\bar{X}^2} [\sum_{i=1}^{n} (y_i - \hat{R} x_i)^2/(n-1)] \qquad (4.18)$$

When interest is in the mean or total of Y, the estimates are as given before: and variances can be calculated from Equation (4.17) by recalling the rule that $V(aR) = a^2V(R)$, where the constant a is now either $\bar{X}$ or $N\bar{X}$, since both of these quantities are assumed known, and thus play the part of constants. Calculation of an estimate of $V(Y_T)$ is easier in the following equivalent form:

$$s^2(\hat{Y}_R) \doteq \frac{N(N-n)}{n(n-1)} [\Sigma y_i^2 + \hat{R}\, \Sigma x_i^2 -2 \hat{R}\, \Sigma y_i x_i] \qquad (4.19)$$

Note that this is the variance for estimating a total.

Since it is advisable to check that the coefficients of variation of the means of Y and X are less than 0.10, another form for calculation of variability is:

$$[c.v.(\hat{Y}_R)]^2 = \frac{s^2(\hat{Y}_R)}{\hat{Y}_R^{\,2}} = \frac{1-f}{n} [c_{yy} + c_{xx} -2c_{yx}] \qquad (4.20)$$

where c_{xx}, c_{yy}, and c_{yx} are the coefficients of variation of y, x and the analogously defined cross-product term:

$$c_{yx} = \frac{\Sigma yx - n\bar{y}\bar{x}}{(n-1)\bar{y}\bar{x}}$$

Readers who refer to Cochran (1977) should note that he used the coefficient of variation of the mean, e.g., c_{yy}/n.

The squared coefficient of variation of Y_R is often termed the relative variance, and can be used to calculate variances of any of the three estimates of interest (the coefficient of variation, being a relative quantity, has the same value for $\hat{Y}_R$, $\hat{R}$, or $\hat{R}_T$).

Confidence limits can be obtained as before:

$$\hat{Y}_T \pm zs(\hat{Y}_T) \quad \text{or} \quad \hat{R} \pm zs(\hat{R})$$

Example 4.6 Ratio corrections for variable plot size

A numerical example of corrections for different lengths of a strip-transect is given by Norton-Griffiths (1975). The data are those from an aerial survey for several species of African "game". Only wildebeest are considered here. The data are as follows:

	x_i	y_i
Transect	Area (km^2)	No. counted
1	8.2	58
2	13.7	44
3	25.8	175
4	25.2	141
5	21.9	151
6	20.9	144
7	23.0	131
8	19.2	135
9	21.4	104
10	17.5	111
11	19.2	130
12	20.8	136
Totals	236.8	1460

The total area of the study region was 2829 km^2, so the population estimate is:

$$\hat{Y}_R = \frac{\Sigma y_i}{\Sigma x_i} X_T = \frac{1460}{236.8} 2{,}829 = 17{,}440 \text{ wildebeest.}$$

There were 126 possible strips in the area, so that N = 126, n = 12, and calculations from Eq (4.19) are:

$$s^2(\hat{Y}_R) \doteq \frac{N(N-n)}{n(n-1)} [\Sigma y_i^2 + \hat{R} \Sigma x_i^2 -2 \hat{R} \Sigma y_i x_i]$$

$= [126(114)/12(11)][193{,}262 + (6.16)^2\ 4{,}935 - 2(6.16)30{,}561]$
$= 436{,}580.$

The standard error is $(436{,}580)^{1/2}$ = 66l, which is quite small compared to the estimate.

It may be noted that the sample size (12) is a good deal less than the 30 recommended as a rule of thumb for using ratio estimation. However, this very likely is an instance where the ratio estimate is nearly optimal, i.e., the relationship goes through the origin, and the variance of the counts likely increases with the area of the transects. Hence, it seems quite reasonable to neglect the n = 30 rule. Density per unit area is estimated by:

$\hat{R}$ = 1460/236.8 = 6.16 wildebeest per km^2.

4.14 Double sampling

The major problem with ratio estimation in ecological studies is just that there are various situations where the method is potentially useful, but a total for the auxiliary variable is not known exactly. Many of these situations do not seem to fit neatly into the present methodology of survey sampling, but it does seem that double sampling comes close enough to provide a useful framework for examining the problems and a useful starting place for much-needed research. The basic idea is just that of the ratio estimation scheme. We have a random variable of primary interest (Y) and an auxiliary variable (X) known to be well-correlated with Y. The missing item is a known total for this auxiliary variable (X_T).

In the instances of interest here, measurements of the auxiliary variable (X) are either very inexpensive to obtain, or are readily available for a large sample taken over the study region. A convenient example is that used to describe ratio estimation; the use of strip transects. We now suppose that the total area of the region under study is not known. If the area is mapped, then it is obviously an inexpensive process to make a large number of measurements of the lengths of potential transect lines from the map. One can thus come very close to estimating the total area (X_T) by working with the map. If we denote this estimated total as X'_T, then double sampling proceeds in just the same manner as ratio estimation, i.e.,

$$\hat{Y}_R = \frac{\Sigma y_i}{\Sigma x_i} X'_T$$

but it is now necessary to make some allowance in variance estimation for the fact that the total(X'_T) of the auxiliary variable is not known exactly. Eberhardt and Simmons (1987) conducted some monte carlo studies to suggest when double sampling might still be useful under this limitation.

If the study region is mapped, there are usually better ways to measure the total area (e.g., by planimetry). However, various nontrivial examples can be considered. The survey may be concerned only with a particular cover type, which is not mapped. If the work on the actual sample transect is quite time-consuming, then it may be well worthwhile to measure only the width of the cover-type on a large number of "auxiliary" transects. These widths then provide an estimate of X_T.

Another situation where double-sampling may be useful is where detailed measurements need to be made on individual plots by some time-consuming process. One example is in estimating total oven-dry biomass of, say, nonwoody vegetation. The time required for clipping, drying, and weighing vegetation severely limits the number of plots that can be so dealt with. Double-sampling might well be utilized by using stem counts as an auxiliary variable, since this can be done on a rather large sample of plots at low cost. A similar prospect exists when chemical analyses are to be done on vegetation, but in this case it may be desirable to use weights on a large sample of plots as the auxiliary variable.

An essential feature of these examples is that accurate measurements of the auxiliary variable can be made in each instance. This appears to be the basis for the present theory of double sampling as given, for example, by Cochran (1977: Ch.12). Unfortunately, there are a great many very useful potential applications in ecological studies that do not seem to quite "fit" the existing theory. These are situations where the auxiliary variable is an estimate of some kind, and is subject to either sampling error, bias, or both. The biomass of vegetation example provides a convenient case. Rather than stem counts, the investigator may choose to use an ocular estimate of biomass on a large sample of plots as an auxiliary variable. With some experience (best gained by guessing weights on a sample of plots and then clipping and weighing), he may become very proficient at visual estimation. The problem now is that the auxiliary variable is subject both to the "chance" errors inherent in visual estimation and to any persistent tendency to consistently overestimate or underestimate.

Another illustration may be taken from aerial censusing of animals. Practically all of the available experience shows that aerial observers tend to miss a substantial fraction of the animals on a sample unit (very often a strip transect). Nonetheless, since aerial surveys can be relatively inexpensive, efforts may be made to "calibrate" the surveys by using some accurate method to enumerate the number of animals actually on a subsample of the plots surveyed from the air. If it can be supposed that these "ground-truth" counts are truly without error, then it can be argued that the requirements of double-sampling are met. The aerial survey now provides the auxiliary variable (X), while the ground count provides the accurate census (Y) that is wanted. However, the auxiliary variable (aerial count) is clearly going to be subject to sampling errors, due to a large variety of causes. Hence we no longer have quite the same situation as when the auxiliary variable can be measured without error. It may be feasible to completely survey the study area from the air. However, this is still not a known total, as a repeat survey flown under identical conditions will without much question yield a different total count.

Many readers will have recognized another problem that was passed by above. This is that the "accurate" measurement (Y) is seldom achievable in census work. Usually the best that can be managed is an estimate that is believed to be unbiased, but is clearly subject to sampling error. We thus have both Y and X subject to sampling errors. This circumstance may bring in some major problems in statistical analysis. These problems are particularly difficult in regression analysis, and remain unresolved for a number of circumstances of importance to ecologists and biologists. Ricker (1973) reviewed the situation for problems in fisheries research and management.

There is thus a need to exercise some caution in applying double sampling in situations where the auxiliary variable is subject to sampling errors, particularly when regression equations are used. In many practical applications in ecology it seems that the ratio approach may be quite satisfactory so we will usually rely on it here.

If it is clear that the relationship does not pass through the origin and if the variance appears relatively constant around a regression line, then it is likely that the regression method should be used. However, in the many cases where it is necessary to assume sampling errors in the auxiliary variable, the usual elementary textbook test for significance of the intercept cannot be trusted. Hence it may be best to depend on judgements as to the nature of the relationship and the pattern of variability in choosing between ratio and regression methods.

4.15 Cluster sampling and subsampling

Cluster samples are likely to be useful in field studies whenever the item of interest is primarily associated with some natural sampling unit. An example might be some species of insect found only on a particular species of plant. Any interest in enumerating the insects, or in studying some other measurement, such as the percent of insects parasitized, requires attention to the fact that they come in clusters. In point of fact, this distinction is often ignored in practice, and it can be safely said that measures of variability obtained without considering the clustering effect will usually be very seriously underestimated. Of course, in the example here described one might reasonably use a ratio estimate, counting the number of plants and sampling some part of them for insect abundance.

In some cases it is possible to deal with clusters that are all comprised of the same number of individual sampling units. This is a natural way to approach large-scale area samples, where the "primary sampling unit" may be taken to be a square mile (section). One may want to use much smaller plots (square-meter or 0.01 ha, perhaps) for the actual measurements, but to enumerate the variable of interest on several such plots in each square mile in the sample. One approach is then to draw a random sample of n square miles from the overall area, and to locate m plots (the subsample) in each of the selected primary units. This is usually termed two-stage sampling. An important consideration in such schemes is determining how many subsamples (m) and the total number (n) of primary units to take to minimize the overall variance (or maximize precision) for a fixed over-all cost.

We will not attempt to detail the procedures for optimum use of subsampling methods, but it is worth mentioning one scheme for calculating the overall variance of an estimate, and thus confidence limits. This is just to use the subsample results for each primary sampling unit to estimate a total for that unit. That is, if there are m plots in each unit, one just obtains the total for those m plots and multiplies it by the reciprocal of the sampling fraction to get an estimate for the primary unit. The primary unit totals can then be used directly as random variables to compute a variance for the survey total. This variance will reflect both components of variability -- that for subsampling (within primary units) and that for differences among primary sampling units. What one loses, of course, is any information on the optimum subsampling rate. The same scheme can be used when the clusters (primary units) are of different sizes (i.e., contain different numbers of sampling units, as, for example trees in woodlots). The shortcoming here is that if the clusters vary considerably in size, that difference will contribute greatly to the overall variance. This point is most important if there is additional information on the cluster sizes in the population, but as was mentioned above, one might then be able to use ratio methods.

Subsampling schemes are often conveniently used to combine systematic and random sampling. In the example above of randomly selected sections (square miles) which are then subsampled, it is often possible to reduce the labor involved if the primary units (sections) are selected at random (usually in a stratified sampling plan) and a series of plots located systematically along a transect within each section as subsamples. It is highly desirable that the transect starting points be randomly selected to avoid any bias due to edge effects or such things as old fencelines in the sections.

Subsampling schemes can involve several stages, and various complexities of estimation. One might for example, use a stratified random sample of square miles, locate subsampling plots in each randomly drawn section, and then elect to examine only a random sample of individual plants on each plot for the variable of interest, which might in turn involve measurements subject to error. Obviously, the statistical analysis of such data can be quite complicated. One way to simplify matters a great deal is to resort to jackknifing or bootstrapping.

Sampling in two (or more) stages is also worth considering when there is uncertainty about the accuracy of the method for making measurements, as is so often the case in estimating the abundance of animal populations. It is usually the case that population density will vary considerably over large areas, and the investigator may have a reasonably good notion of how density varies with habitat and so on (or this may be a major item of interest). It is then logical to use a stratified random sampling scheme to locate primary sampling units on which the actual measurements of density will be attempted. This does not, of course, reduce any uncertainty in the actual measurement method, but it does keep the area differences from compounding matters.

Example 4.7 A cluster sampling example

One simple example of cluster sampling was mentioned in Example 4.5 (stratified sampling). The "primary sampling units" (square miles) were selected at random, and then subsampled with a cluster of eight small plots. All that is needed for analysis of the resulting data is just to multiply the total for the eight subsamples by a "raising factor" or "blow-up factor", which is simply the reciprocal of the sampling rate. In the example used, the individual plots were each 1/50 of an acre, hence the necessary adjustment factor is: 640/(8/50) = 4000. Once this is done, the remaining analysis proceeds as though no subsampling had taken place. Skeptics may need to do a little algebraic manipulation at this point. When subsampling rates are not constant, things become somewhat more complicated, and a sample survey text should be consulted for details. However, if the subsampling rate does not vary greatly, the same procedures can be used without elaboration. All that happens is that one overestimates the variance, in most situations. But if the subsampling rate varies considerably and/or is related to size of the primary sampling unit, then by all means consult a textbook on sampling or a statistician.

Supposing constant size of the primary sampling unit, and a constant subsampling unit (the case most likely in ordinary applications) the main question to be settled is "What is the best subsampling rate?". As usual, answers depend on relative costs. That is, a particular effort (hence cost) is required to survey an individual sampling unit (i.e., one plot in the example), while a separate cost is engendered by the time and travel going from one primary sampling unit to the next. For a given total expenditure for the entire survey, the optimum subsampling rate is that which minimizes the overall variance given the above two costs

Since natural populations exhibit a somewhat frustrating tendency for variances to change nonlinearly with size of the sampling unit (plot size), a simple equation for subsampling rate is not available. What's really needed is a "variance law", i.e., a relationship between plot size and variance. To obtain such a relationship, one has to run a special study using several plot sizes. Then it becomes possible to incorporate costs and get on with the business at hand by consulting Cochran (1977, Ch. 9). As we noted earlier, the kind of measurement (weights, counts, etc.) and the organism under study influence the "variance law" substantially. Hence there are two choices open at this point. One is to run a fairly expensive preliminary field study, and thus to manufacture your own "variance law". The second choice is to resort to the literature in the appropriate field, seeking papers in which several different plot sizes have been used. A number of references along these lines appear in Eberhardt (1978a). However, it is clear that this is an area needing rather more research attention in ecology.

Example 4.8 Cluster sampling involving proportions

One of the commonest errors in the ecological literature is an uncritical acceptance of the binomial distribution as an appropriate model for analysis of proportions in data collected in clusters. It is the appropriate model if, and only if, a simple random sample of individuals can be obtained. In practical problems one almost always collects observations as clusters. When this is the case, the clustering effect must be taken into account in order to obtain a meaningful variance. Very rarely do we encounter a population so well mixed that clusters are indeed equivalent to simple random samples, so that such an example is likelyto be more of a curiosity than anything else.

The simplest way to deal with cluster sampling for proportions is to treat the individual observations as random variables. In this instance, the appropriate form of the ratio estimator is:

$$\bar{p} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i}$$

where y_i denotes the number of individuals in the i^{th} cluster possessing the attribute of interest, and x_i is the total number of individuals in the i^{th} cluster, while n is the number of clusters.

The appropriate variance estimate here is (Cochran 1977:65):

$$V(\bar{p}\ ) = \frac{\Sigma(p_i-\bar{p})^2}{n(n-1)}$$

where $p_i = y_i/x_i$, i.e., the observed proportion in the i^{th} cluster. (We here neglect the finite population correction which can be inserted as a multiplier (1-f) if needed).

An interesting set of data to illustrate behavior of proportions in clusters comes from a paper by Johnson and Chapman (l968). This was a study to estimate the number of fur seal pups on a "rookery" on the Pribilof Islands, off Alaska. A large sample (4,965) of pups were marked (in groups) and then clusters of100 were examined (for the proportion marked) at randomly selected sampling stations. The estimate of the total number of pups on the rookery was obtained from

$$\hat{N}_1 = \frac{M}{(\bar{p})}$$

where N is the population estimate, M is the number marked (4,965) and $\bar{p}$ is the mean proportion marked, calculated as in the above example.

Two ways of estimating the variance were used. One is based on the "delta method", and is:

$$V(\hat{N}_1) = \frac{M^2 v(\bar{p})}{\bar{p}^4}$$

where $v(\bar{p}\ )$ is obtained as in the above example. The second method is that of "interpenetrating" sampling, in which the sample is subdivided randomly into a number of subsamples. A separate estimate of the population size $\hat{N}_i$ is made from each subsample and these are then averaged for the final estimate, i.e.:

$$\hat{N}_2 = \frac{1}{r}\sum_{i=1}^{r}\hat{N}_i$$

and:

$$v(\hat{N}_2) = \sum\frac{(\hat{N}_i\ -\ \bar{N})^2}{r(r-1)}$$

It should be noted that the two estimates of the total population will not necessarily be identical, nor will the variance estimates be the same for the two methods.

The observed numbers of marked pups in clusters of 100 (recorded on two sampling dates) were:

August 26, 1961 (25 samples) 2, 0, 1, 6, 4, 33, 62, 49, 55, 38, 52, 77, 85, 54, 27, 17, 3, 3, 3, 2, 2, 1, 0, 0, 4.

August 28, 1961 (58 samples) 0, 0, 0, 4, 0, 0, 0, 12, 4, 8, 60, 48, 72, 72, 76, 80, 56, 44, 50, 56, 56, 28, 60, 36,44, 44, 28, 52, 72, 28, 72, 60, 60, 84, 76, 52, 84, 48, 52, 60, 40, 12, 8, 12, 4, 8, 44, 16, 0, 8, 0, 0, 4, 12, 8, 0, 0, 0.

The interpenetrating or replicated samples were defined as follows:

Subsamples 1, 2, 3: Every third observation of August 26, beginning with observations 1, 2, 3, respectively.

Subsamples 4-10: Every seventh observation beginning with observations 1, 2, 3, 4, 5, 6, 7, respectively.

Since there were 25 observations on 26 August, this procedure yields subsamples of size 9, 8, and 8, respectively, while the 58 observations on 28 August yield two sets of size 9 and 5 of size 8. These data lead to the following estimates for the interpenetrating sampling:

Subsample	N_i
1	20,497
2	24,219
3	20,060
4	17,455
5	16,674
6	17,732
7	14,391
8	12,490
9	13,066
10	14,821
Total	171,405

Averaging gives $\hat{N}_2 = 17{,}140$ with $v(\hat{N}_2) = 1{,}353{,}000$, while $\hat{N}_1 = 16{,}550$ with $v(\hat{N}_1) = 2{,}950{,}000$.

4.16. Some additional sampling techniques

There are a number of additional techniques students should know about. Multistage sampling was used in Examples 4.4 and 4.5 where subsamples of the primary sampling units were actually enumerated. As pointed out there, it isn't necessary to consider the subsampling in obtaining a variance estimate. All that is needed is to use the subsample data to make estimates for the primary sampling units and treat those values exactly as one would if the entire unit had been tallied. However, it may be desirable to consider the “within sampling unit” variability in order to do a more efficient job of designing the survey. This requires more complex equations which are given in many books on sampling [e.g., Cochran (1977), Thompson (1992)].

Another useful technique uses unequal probabilities in selecting samples. This approach is exemplified by the line intercept technique described in Chapter 5, and may be useful in any circumstance where the probability of selection may vary from unit to unit, either naturally or for convenience or improved efficiency. Texts or references to sampling techniques may refer to the Hansen-Hurwitz estimator. This is a method for using unequal sampling probabilities (see any of the sampling texts for details).

A relatively recent development is known as adaptive sampling. This may be a very useful approach when items of interest tend to be clustered, but in such a manner that there is no readily defined unit that contains all of the elements of a cluster. The technique provides a means for expanding the sampled area around primary units where a concentration of the items of interest is encountered, without biasing the results (which occurs with certainty if one simply expands the area to include more individuals). Details appear in Part IV of Thompson (1992) and a more extensive (and more theoretical) treatment appears in Thompson and Seber (1996).

Another potentially valuable approach is generally known as “kriging” after the South African mining engineer, Krige, who developed the initial approach in searching for profitable sites for mining for gold or other minerals. The approach is now used in petroleum exploration. In both of these examples drilling exploratory holes can be very expensive and time-consuming. The methodology thus utilizes spatial correlations among the available samples to estimate abundance or density on an area. A natural descriptive phrase thus is “spatial sampling”, and there are many instances where this may be useful in ecology. Thompson (1992:Part V) gives a useful summary and references to the extensive literature.

4.17 Exercises

4.17.1 Using a table of random numbers

Drawing a sample with the aid of a table of random numbers is not very complicated, but the student should try drawing a sample of 10 individuals from a population of 20, and another sample of 10 from a population of 1000 (the "populations" can be just the numbers 1-20, and 1-1000). Two approaches to starting points in the table may be considered. One is to somehow make a "random" start, (e.g., by closing one's eyes and touching a point on a page to select random coordinates in the table for a starting point) the other is to mark off sets of digits as they are used, going on through the table as different occasions for its use come up. The latter course is preferable for repeated surveys of the same areas. Note that samples of 10 out of a small population (like 20) may yield one or more repetitions of random numbers. Notice, too, that one has to use a two digit column of numbers, and many must be rejected with a population of 20. This seems to be even more of a problem with the population of 1000, since one should use 4 digits in order to permit the number

1000 to have a chance to be drawn. However, it is simple to arbitrarily assign the number 1000 to the 3 digit sequence 000 and thus use three columns (001 to 999, plus 000 for 1000). When working with EXCEL it is convenient to use the RANDBETWEEN() function, as that avoids the need to use a table of random numbers.

4.17.2 Determining sample size

Suppose that we want 95% confidence limits of about ± 15% for the data in Example 4.1. What sample size is required if N =1000? Calculate sample size for ±10% for N = 1000.

4.17.3 An exercise in allocation

As an exercise in allocation, use the values of s_h^2 actually obtained in the caribou survey in Example 4.5 to calculate a new allocation and compare it with that actually used.

Another way to guess at the S_h to use for allocation is to assume the coefficient of variation $(s/\bar{x})$ is constant. Calculate the c.v.'s for each stratum, and try a "typical" values for allocations. Are there substantial differences between the various schemes? Comment on the results.

4.17.4 Computations for mortality survey

Compute $\bar{y}_{st}$ and the total mortality estimate for Example 4.4 along with 95% confidence limits. It is often convenient to use $2[V(\bar{y}_{st})]^{1/2}/\bar{y}_{st}$ as "percentage limits"on survey results. Compute that value and compare it with the same result for example 4.3.

4.17.5 Stratified sampling in a vegetation study

A survey designed to estimate biomass of non-woody vegetation in a sagebrush stand (Eberhardt and Rickard 1963) provides an example of a different approach to stratification and illustrates some of the potential flexibility of sampling methods. In this example, proportional allocation was used in order to avoid advance preparations other than marking out the area well enough to avoid recounting individual plants. Two investigators worked together, one classifying and tallying each sagebrush plant into one of five strata, while the other checked off each plant on graph paper on which certain squares had previously been randomly selected as representing a plant to be sampled (it was thought that about 1/30 of the bushes should be sampled, so three numbers from 1-90 were designated as meaning"sample" and a table of random numbers was used to produce the sampling chart). When a "winner" turned up, the bush was subdivided into from two to five parts, and one of the parts was randomly selected. If that part was too large for weighing, it was subdivided and a random selection again made. The selected portion was than clipped, oven-dried and weighed.

Data from the survey are tabulated below. The "subsampling fractions" show approximately how much of an individual plant was actually removed. In the first record in the table, about 1/4 was removed, while in the second case there were two sub-divisions, and roughly 1/10 [(1/5) times (1/2)] was actually removed. Thus to estimate total weight for a given plant one would multiply by 4 or 10. Stratum IV contained the largest plants and the two plants actually sampled were sampled at rates of 1/56 and 1/28, respectively. Of course the divisions were not exact but any errors in subdividing will enter into overall variance of the survey estimates. There was actually a fifth stratum, but only one plant was sampled, so it has been left out of the tabulation.

As an exercise, the student should work out an estimate of mean oven-dry material and its variance for the entire sagebrush stand using the data in Table 3.l. Calculate an optimum allocation for a sample of the size used here (25), and compare with the proportional allocation (neglect the fpc). Calculate coefficients of variation. Comment on the results.

Results of stratified sampling of a sagebrush stand.

Stratum	Number of bushes in stratum	Subsampling fractions for sampled bushes	Oven-dry weight of sample (g)
I	169	1/4	0.60
		1/5,1/2	1.90
		1/2,1/2	2.05
		1/2,1/2	1.05
		1/2	1.20
II	309	1/4,1/2	1.60
		1/5,1/4	3.20
		1/5,1/5	1.45
		1/3,1/3	4.05
		1/3	2.05
		1/3,1/2	1.45
		1/4,1/4	2.40
		1/3,1/2	1.65
		1/3,1/4	0.60
III	301	1/5,1/4	1.85
		1/5,1/2	2.85
		1/5,1/4	7.15
		1/5,1/3	2.15
		1/3,1/4	4.10
		1/4,1/5	3.50
		1/4,1/3	5.25
		1/5,1/3	5.60
		1/3,1/3	1.55
IV	57	1/8,1/7	6.05
		1/7,1/4	3.60

4.17.6 Try jackknifing to calculate a standard error for Example 4.6. Compare your result with that given in the Example (661). Also, calculate bias estimates for $\hat{R}$ using jackknifing and bootstrapping. Use 200 bootstraps. There is only one jackknife estimate of bias available, but you can run the bootstrap repeatedly and see how the bias changes. Comment on your results. Don't forget to consider the magnitude of the bias relative to the estimate.

4.17.7 Bootstrap the data for August 26 (n=25) from Example 4.8 and compare your results with the ratio estimators N-hat(2) and V(N-hat(2)) given in the example, and with the binomial variance estimate given below. Do 200 bootstraps and calculate Bias(boot) from eq.(3.2). Run repeatedly and see how Bias(boot) varies. Is there an indication of appreciable bias? Recall that when simple random sampling of individuals is assumed:

$$v(p) = pq/(n-1)$$

where n here is 2500. The difference in the two estimates reveals why the binomial formula should never be used with cluster samples.

4.17.8 Try jackknifing the "interpenetrating sampling" results of Example 4.8, and compare the variance you get with that given in the example. Explain the results.